A Novel Methodology for Determining Effectiveness of Preprocessing Methods in Reducing Undesired Spectral Variability in Near Infrared Spectra

Jhon Buendia Garcia, Julien Gornay, Marion Lacoue-Negre, Sílvia Mas Garcia, Jihane Er-Rmyly, Ryad Bendoula, Jean-Michel Roger

A novel analysis methodology for preprocessing methods effectiveness determination in reducing undesired spectral variability in near-infrared spectra acquisition

Jhon Buendia Garcia[a,c], Julien Gornay[a], Marion Lacoue-Negre[a], Silvia Mas Garcia[b,c], Jihane Er-Rmyly[a], Ryad Bendoula[b], Jean-Michel Roger[b,c]

[a] IFP Energies Nouvelles, Rond Point de l'échangeur de Solaize, France
[b] ITAP-INRAE, Institut Agro, University Montpellier, Montpellier, France
[c] ChemHouse Research Group, Montpellier, France

Corresponding Authors:
Jhon Buendia (jhon.andersson@gmail.com)

Abstract

This study uses a novel analysis methodology based on the Hierarchical Clustering Analysis (HCA) to determine the effectiveness of different preprocessing methods in minimizing undesired spectral variability in near-infrared spectroscopy due to both the consecutive and repetitive acquisition of the spectrum and the sample temperature. Nine preprocessing methods and different combinations of them were evaluated in four case studies: reproducibility, repeatability, sample temperature, and combination of the before mentioned cases. Eighty-four spectra acquired on seven different hydrocarbon samples from catalytic conversion processes have been selected as the real case study to illustrate the potential of the mentioned methodology. The approach proposed allows a more detailed discriminatory analysis compared to the classical methods for comparing the between-class and the within-class variances, such as the Wilks' lambda criterion, and hence constitutes a powerful tool to determine adequate spectral preprocessing strategies. This study also proves the potential of the discrimination analysis methodology as a general scheme to identify atypical behaviors either in the spectrum acquisition or in the measured samples.

Keywords

Spectral variability, Hierarchical Clustering Analysis (HCA), Principal Components Analysis (PCA), Preprocessing effectiveness, outliers, Qresidual, Hotelling's T^2, reproducibility, repeatability, sample temperature.

1 Introduction

In the past few decades, the use of Near-Infrared Spectroscopy (NIRS) in the development of non-destructive and rapid measurement applications has been significantly increasing in several industries, such as food[1, 2], pharmaceuticals[3, 4], and petroleum[5, 6]. Due to the recent growth boom in using the NIR spectrum for real-time acquisition data[7,8], the need to determine, analyze, and minimize spectral variability that is not associated with the physicochemical characteristics of the sample has notably arisen. This need becomes particularly evident when spectral variability is mainly generated by factors associated with the spectrum acquisition, such as spectrometer system,

operator, measurement conditions, and environmental factors such as temperature and humidity, rather than the physicochemical characteristics of the sample. An example of this is the possible generation of spectral variability in the consecutive and repetitive acquisition of NIR spectra on a sample whose physicochemical characteristics remain constant over the spectrum acquisition. A lack of minimization of this type of spectral variability can result in inaccurate analysis and interpretation of spectroscopic information, misleading conclusions and flawed decision making[9, 10].

Among the classical performance parameters needed to validate a measurement methodology, precision is the most affected by the aforementioned factors. Precision is defined as the closeness of agreement between measured values obtained by replicate measurements on the same or similar samples under conditions of repeatability or reproducibility[11]. Repeatability conditions include the same measurement procedure, the same operator, the same instrument and measurement conditions, the same location, and a short interval between repetitions[12]. On the other hand, reproducibility implies successive measurements of the same sample under changing measurement conditions[13], such as measurement principle, measurement method, operator, measurement instrument, reference standard, location, conditions of use, and time. The NIR spectra acquisition is very sensitive to any change in measurement; even ensuring that both the spectrum acquisition conditions and the physicochemical characteristics of the sample do not change in a repetitive NIR spectrum acquisition, the resulting spectra may have differences that can lead to random errors and deviations, which must be corrected or minimized.

Due to its high impact on NIR spectra acquisition accuracy, the temperature is the most studied influencing parameter[14, 15]. Hansen *et al.*[16] showed that molecular bonds vibration intensity depends on temperature, leading to changes in the spectrum according to temperature variation. Furthermore, some physicochemical properties of samples, such as viscosity and density, are temperature-dependent, and many changes in the sample due to temperature are not permanent and do not reflect the intrinsic nature of the sample[17–19]. Nevertheless, these changes can significantly affect spectrum acquisition. As with the spectral variability generated by repetitive spectrum acquisition of a specific sample, the variability caused by sample temperature must be minimized to ensure the reliable description of the sample physicochemical behavior from the spectroscopic information extracted.

Data preprocessing is a common step for reducing undesired effects and for minimizing spectral variability. There are different preprocessing algorithms for the correction of the undesired spectral variation; these can be divided into two main categories: scatter-correction methods, employed to correct the additive and multiplicative effects, and spectral derivatives, used to minimize the sources

of unwanted and non-informative spectral variations[20]. Among the most commonly preprocessing methods used in NIR spectroscopy, Savitsky-Golay derivative (Sav-Gol)[21], Extended Multiple Signal Correction (EMSC)[22], Standard Normal Variate (SNV)[23], and recently, the Variable Sorting for Normalization (VSN)[24], can be highlighted. However, the effectiveness of the preprocessing methods is highly dependent on the type of spectroscopic information analyzed and the factors that are causing its variability[9].

The preprocessing method effectiveness evaluation is generally based on the performance of prediction models[25–27]. Among the contributions reported in the literature, the work of Gerretzen et al.,[28] which presents a novel approach for the selection of the most appropriate preprocessing methods based on the design of experiments, is worth mentioning. Similarly, the studies of Devos et al.,[29] and Allegrini et al.,[30] which, by means of a parallel workflow approach of preprocessing and variable selection, present an interesting alternative to the optimization of the preprocessing method selection. Nonetheless, the application of these approaches may be limited when the variability of the physicochemical characteristics of the samples is negligible, but significant spectral variability exists as a result of the repetitive spectrum acquisition and the sample temperature. In that case, a different analysis approach may yield more detailed results, helping to improve understanding of the impact of these parameters. Another less common approach to assessing preprocessing methods effectiveness is analyzing the spectral variance[31]. Different statistical tools are available to determine both within-class variance (multiple measurements of the same sample) and between-class variance (measurements of different samples). One of the most common criteria used to evaluate between-class and within-class variances is the Wilks' lambda[32].

In this study, a novel and general strategy based on the Hierarchical Clustering Analysis (HCA)[33] was proposed for evaluating the effectiveness of preprocessing methods in reducing the spectral variability generated by parameters related to the continuous and dynamic spectrum acquisition. To this aim, the effectiveness of nine preprocessing methods and different combinations of them in minimizing undesired spectral variability due to repeatability, reproducibility, sample temperature and combination of these parameters were evaluated. Eighty-four spectra acquired on seven different hydrocarbon samples from catalytic conversion processes have been selected as the real case study to illustrate the potential of the mentioned methodology.

To obtain reliable conclusions and validate the results obtained by the analysis methodology proposed, the Wilks' lambda criterion[32] was used as a reference method.

2 Material and methods

2.1 Samples

Twenty-four vacuum gasoil (VGO) samples were processed in the catalytic conversion pilot plant reactors at IFPEN (Solaize, France). From these reactors, ninety-three different hydrocarbon samples, known as total effluent, were obtained (see references[34, 35] for a detailed description of catalytic conversion processes). From these 93 samples, 7 samples were selected, ensuring their representativeness and physicochemical diversity. Table 1 summarizes four relevant physicochemical properties of the selected samples: the density[36] and the simulated initial boiling point and distillation temperatures range to obtain both 5% and 95% of sample distillate (Simulated Distillation IBP, T5 and T95)[37]. It can be observed that physicochemical variability between the selected samples is guaranteed.

Table 1 Samples physicochemical properties. SimDis IBP, T5 & T95 description: Simulated distillation to determine the temperatures to start the sample evaporation and to recover both 5% and 95% of sample distillate

	Sample Physicochemical Properties			
		Distillation Temperatures		
Sample ID	**Density (gr/ml)**	**IBP (°C)**	**SimDis T5 (°C)**	**SimDis T95 (°C)**
Sample 1	0.8049	84.7	121.6	425.8
Sample 2	0.8186	79.8	111.5	474.5
Sample 3	0.8219	80.3	117.8	516
Sample 4	0.8411	83.8	136.1	503.5
Sample 5	0.847	83.7	129.1	512.6
Sample 6	0.8962	148.9	227.6	504.6
Sample 7	0.9181	159.8	231.6	502.9

2.2 Spectral acquisition

The spectra were recorded with a Fourier Transform Near-Infrared spectrometer (FT-NIR) MATRIX-F (Bruker, Optik GmbH, Ettligen - Germany) within the range of 9090 - 4600 cm^{-1} (1100 - 2160 nm) and a resolution of 4 cm^{-1}. 32 scans were used to obtain the final spectrum after each measurement. For acquiring absorbance spectra, the spectrometer system was equipped with an immersion transflectance probe with an optical path fixed at 2 mm withstanding temperatures ranging from -40 °C to 200 °C. The software used with the spectrometer was OVP (OPUS Validation Program - Bruker, Optik GmbH, Ettligen - Germany) which automatically performs a series of analyses of the instrument's performance, evaluates them and ensures that it is operating within specifications. Besides, to ensure the spectrometer operation within specifications and that the spectral variability generated was due to the parameters evaluated and not to the instrument's inadequate functioning, the spectrometer performance was validated once a day using cyclohexane as an external reference

sample. Before NIR analysis, the samples were heated in closed flasks at 60°C for 1 hour in a water bath and shaken manually to ensure their liquid state and homogeneity. The initial boiling point (IBP) reported in Table 1 guarantees no loss of volatiles.

Ensuring the integrity and stability of both the sample and the NIR spectrum acquisition conditions, spectral variability due to repeatability, reproducibility, and sample temperature was generated. A short description of the spectrum acquisition for the cases evaluated in this study is presented below.

- **Case 1. Spectral variability due to reproducibility:** Each of the seven samples was analyzed one time per day during five consecutive days at 60 °C. Thirty-five spectra were obtained.
- **Case 2. Spectral variability due to repeatability:** Each of the seven samples was analyzed three times on the same day at 60 °C. All samples were analyzed in less than 8 hours. Twenty-one spectra were obtained.
- **Case 3. Spectral variability due to the sample temperature:** Each of the seven samples was analyzed at five different temperatures, ranging from 60 °C to 80 °C with a temperature increment of 5°C. The samples were heated in closed flasks at the desired temperature for 1 hour. Evaporation losses of volatiles were null or negligible (see IBP in Table 1). Twenty-eight spectra were obtained.
- **Case 4. Spectral variability due to the combination of the aforementioned cases:** In this case, all spectra acquired in the above-described cases were used.

For each case, a matrix was generated. Each analyzed sample was defined as a class; thus, seven classes were defined in all matrices.

2.3 Analysis methodology

The main steps of the data analysis workflow proposed in this study are schematized in Figure 1. A brief explanation of the procedure is given as follows.

The first step consisted in preprocessing each of the generated matrices. The nine most commonly preprocessing methods used in NIR data were divided into two categories; the filtering and normalization methods. The preprocessing methods from each category were analyzed individually. If the total reduction or compensation of the studied spectral variability were not achieved, the evaluated preprocessing method were complemented with the methods belonging to the opposite category. This allowed the evaluation of all possible combinations and order of use of the preprocessing methods. The methods evaluated are described in Table 2. It should be emphasized that each preprocessing scenario evaluated includes the data centering by columns.

Table 2 Preprocessing methods description

#	Category	Name	Acronym	Parameters
1	Filtering	Automatic Weighted Least Squares Baseline[20]	AWLS-B	
2		Norris-Williams Derivation[38]	NW-D	15-point window, gap size = 7, First order derivation
3		Savitsky-Golay Derivative[21]	SG-D	15-point window, polynomial order = 2 First order derivative
4		Detrend[23]	Dtd	Polynomial order = 1
5		Extended Multiplicative Scatter/Signal Correction[22]	EMSC	Reference spectrum (basis to remove the scatter) = mean of each matrix generated, polynomial order = 2, whole spectral range
6	Normalization	Multiplicative Signal Correction[39]	MSC	Reference data = mean of data, whole spectral range
7		Standard Normal Variate[23]	SNV	
8		Probabilistic Quotient Normalization[40]	PQN	
9		Variable Sorting for Normalization[24]	VSN	Automatic calculation

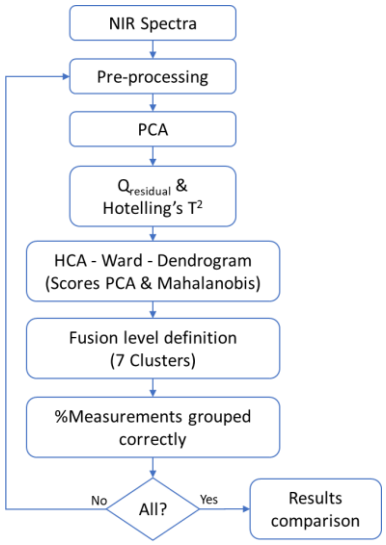


Figure 1 – Methodology flow diagram

Afterward, a preliminary inspection and dimension reduction of the corresponding dataset was performed by principal component analysis (PCA)[41]. The number of the chosen principal components (PCs) captured at least 99 % of the total variance in the dataset. The Q residual and Hotelling's T^2 tests were performed to determine the possible presence of anomalous data[42].

The chosen PCs scores were then used to perform the hierarchical clustering analysis (HCA) employing Ward's algorithm and the Mahalanobis distance. The HCA aims to group clusters to form a

new one to either minimize a statistical distance between classes or maximize a measure of similarity between them[43, 44]. The analysis starts with as many groups as individuals contained in the dataset. From these initial groups, clusters are formed in an ascending manner until all cases treated are included in at least one of them. Ward's algorithm seeks to minimize each group variance by calculating all samples mean in each cluster. The algorithm then calculates each case distance and the cluster mean, adding up the distances between all cases. Finally, the clusters whose sum of distances is minimal are grouped. This procedure creates homogeneous groups of a similar amount of individuals. For achieving the grouping of classes, it is necessary to define a comparison parameter to calculate the variance of each class concerning the others. The most common is the Mahalanobis distance[39].

A widespread manner of displaying the cluster analysis results is constructing a tree diagram known as a dendrogram. The resulting diagram shows the different groups' clustering order and the association measure's value, also known as the fusion level. The fusion level was defined for obtaining seven clusters corresponding to the seven classes. Finally, the number of correctly grouped sample measurements in each cluster was determined, and the percentage of clustering was calculated as: Number of correctly grouped samples / Total number of samples * 100.

These steps were repeated for each preprocessing method scenario, and the results obtained were compared using the percentage of samples correctly grouped as a figure of merit to determine the effectiveness of the preprocessing methods evaluated.

All the analyses were conducted with the PLS_Toolbox version 8.8 (Eigenvector Research Inc., Wenatchee, WA, USA) for MATLAB version R2019b (MathWorks, Natick, MA, USA).

2.4 Results comparison

To validate the results obtained with the methodology proposed in this study, the Wilks' lambda was used as a comparative criterion. This criterion evaluates how well the data set classes are separated by calculating a ratio involving between-class and within-class variances. Several versions of the Wilks' lambda exist. In this article, the ratio of the between-class variance over the total variance was used. This ratio varies between 0 and 1, where 0 means that all the classes are superimposed, and 1 means that all the classes are perfectly separated.

3 Results and discussion

This section shows the results obtained from the application of the proposed methodology. A comprehensive description of its application for the case study of reproducibility (Case 1) is presented. However, only the main results from the case studies of repeatability, temperature effect,

and the combination of all cases are showed. Finally, the validation and the advantages of the proposed methodology are offered.

3.1 Case 1 – Reproducibility

For the reproducibility case, 35 spectra were used (5 spectra for each sample, see section 2.2). Figure 2 shows the raw spectra over the entire spectral range used. At first glance it can be observed the variability of spectra due to the physicochemical nature of the sample, showing, with some exceptions, a trend consistent with the properties reported in Table 1; i.e., the spectra acquired on the sample with the lowest density (sample 1 - dark blue color) is at the bottom of the plot, and the spectra acquired on the sample with the highest density (sample 7 - red color) is at the top. This observation becomes more evident in the black square of Figure 2 where the spectra are magnified over a defined range (1340nm - 1430nm). In this same figure it is possible to visualize that the reproducibility measurements made on a single sample generate a variability that has a similar behavior to the variability caused by multiplicative effects, which can impact the final results obtained.

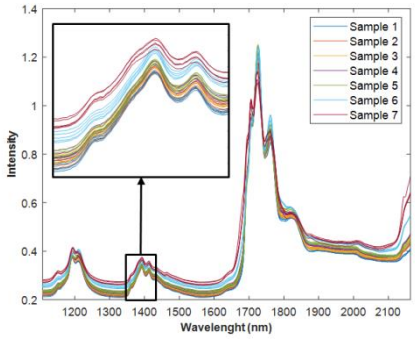
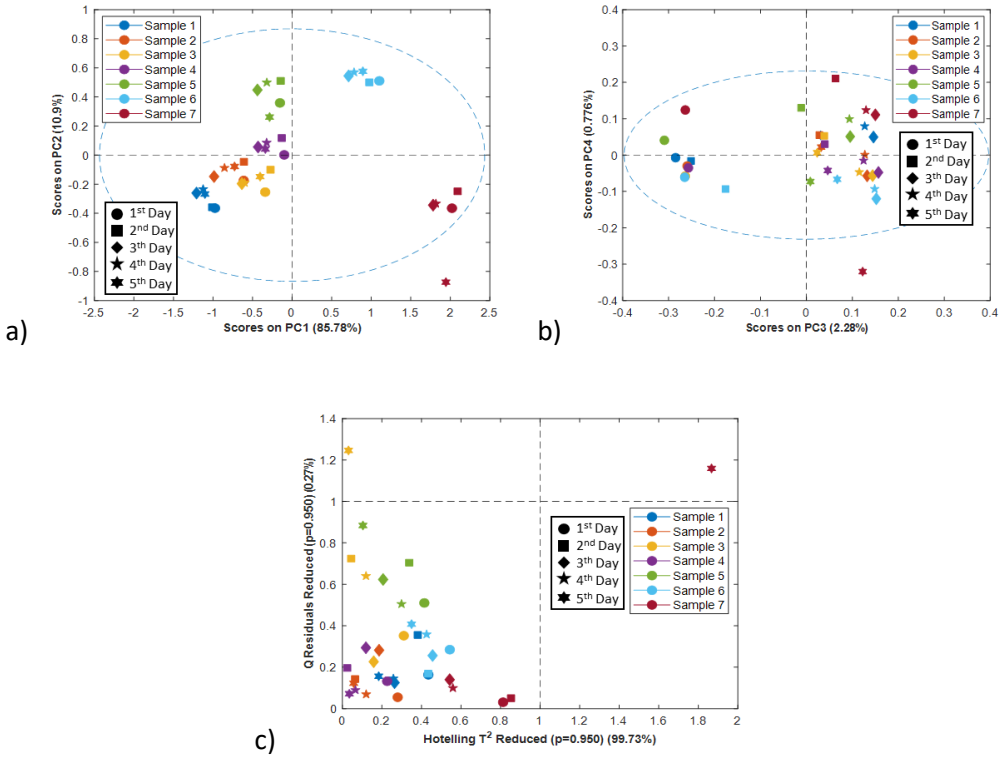


Figure 2 – 35 raw spectra used in Case 1 evaluation over the entire spectral range. Legend: Color → samples analyzed. Black square → 35 raw spectra used in Case 1 evaluation magnified over the 1340nm - 1430nm spectral range

In this first case, the proposed methodology application shows the impact of the variability in NIR spectra acquisition caused by reproducibility measurements as well as the effectiveness of preprocessing methods in minimizing this spectral variability.

Firstly, an exploratory analysis using PCA was performed to visualize the similarity/dissimilarity among the 5 spectra (1 per day) acquired for each of the 7 hydrocarbon samples analyzed. The data set (35 spectra) was merely centered. The first two components (PC1 and PC2) explained 96.7% of the data set total variance. From the score plot of the first two components (see Figure 3a), it can be seen that all 5 measurements of samples 1 to 6 are grouped in a consistent pattern regarding to the variance within classes. However, there are measurements of different samples that intersect with

each other (between-class variance) (see Figure 3a-b), and hence their clustering can be influenced. Additionally, from these score plots, it can also be observed that the measurement of sample 7 on the fifth day is relatively distant from the other measurements of this sample. This could mean the presence of possible outliers in the dataset. Figure 3c shows the Hotelling's T^2 and Qresidual scores for the dataset. It can be observed that the same measurement identified previously (Sample 7, 5th acquisition day) was found to be above the threshold of both tests. Therefore, this measurement can be confirmed as an outlier.

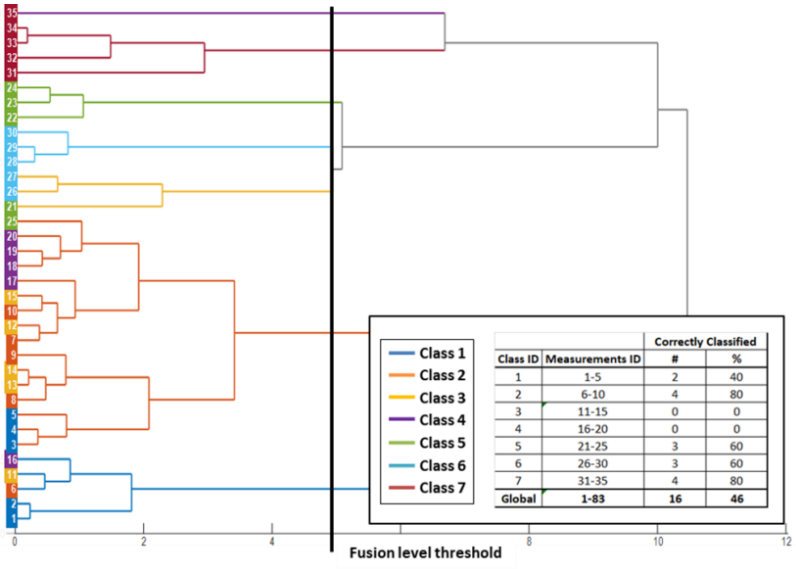


a) b) c)

Figure 3 – a) Score plot of PC1 & PC2 for Case 1 with centered spectra, b) Score plot of PC3 & PC4 for Case 1 with centered spectra, c) Reduced Qresidual & Hotelling's T^2 for Case 1 using 4 PCs with centered spectra. Legend: Shapes → acquisition day. Color → samples analyzed (classes)

The scores of the first 4 principal components yielded by the PCA analysis (99% explained variance) were used to perform the HCA analysis. Figure 4 shows the dendrogram obtained from the HCA, where a fusion level (black line) for obtaining seven clusters corresponding to the seven hydrocarbon samples is defined. This Figure also shows the correct grouping percentage achieved for each class. From the table embedded in the Figure, it can be observed that no sample achieves an accurate grouping of all its measurements. Samples 2 and 7 (orange and red classes) are the classes with the highest correct grouping percentage (4 out of 5 for 80%), while classes 3 and 4 (yellow and purple classes) do not have any correctly grouped measurements. Furthermore, measurement 35 that belongs to class 7 (red class) and identified in the previous steps as a potential outlier is the only

measurement grouped in class 4 (purple class). Based on these results, it can be considered that this measurement has no similarity with any of the other 34 measurements, which confirms its outlier status.



Class ID	Measurements ID	Correctly Classified #	%
1	1-5	2	40
2	6-10	4	80
3	11-15	0	0
4	16-20	0	0
5	21-25	3	60
6	26-30	3	60
7	31-35	4	80
Global	1-83	16	46

Figure 4 – Dendrogram for Case 1 (all measurements) with centered spectra. Legend: Color samples measured (classes). Table: correct grouping percentage for each class

In order to compensate the spectral variability caused by the lack of reproducibility and hence achieve a better clustering of all the classes, the use of appropriate preprocessing methods is needed. Nine preprocessing methods and different combinations of them were evaluated. Figure App 1, which can be found in the article's supplementary section, summarizes the correct grouping percentage results for each preprocessing scenario evaluated in the 4 cases. From this Figure, it can be seen that for Case 1 (diamond shape - blue color), none of the evaluated scenarios reaches a correct grouping percentage of 100%, implying that none of the scenarios achieved the total reduction of the spectral variability generated by the reproducibility measurements. With a correct grouping of 85%, the EMSC was the most effective preprocessing method scenario. Figure 5 shows for this scenario the score plot of the first two principal components of the PCA analysis and the Qresidual & Hotelling's T^2 results achieved using 2 PCs. It can be seen that all measurements are consistently grouped and are within the threshold of the two tests, except for measurement 35 (potential outlier identified). From the HCA dendrogram and the table of its corresponding correct grouping percentage (see Figure 6), it could be observed that this measurement is still wrongly grouped as the unique measurement in class 2 (orange class). From these results, it can be assumed that all other measurements could be correctly grouped without this misgrouped measurement. Therefore, measurement 35 was removed from the data set, and the preprocessing methods scenarios were re-evaluated to confirm this assumption.

Figure App 1 (diamond shape – red color) shows that by removing measurement 35 from the data set, correct grouping of all measurements is possible in 6 scenarios (EMSC, AWLS-B+MSC, AWLS-B+SNV, AWLS-B+VSN, SG-D+SNV, and MSC+AWLS-B). However, only one scenario uses a single method, which is the EMSC. In order to prevent loss of relevant information, the use of a minimum number of methods in the data preprocessing is generally recommended. Therefore, in this case, EMSC could be selected as the most efficient preprocessing method scenario to reduce the spectral variability due to the lack of reproducibility (see HCA dendrogram in Figure App 2 in the supplementary section).

Figure App 1 also shows that both the order of use and the combination of preprocessing methods can influence the final result. An example of this statement can be seen by comparing the scenarios from Case 1 without the measurement 35, where the NW-D method was combined with the MSC, PQN, SNV, and VSN methods. When the NW-D method (filtering category) is used before applying the other preprocessing methods (normalization category), the correct clustering is lower (about 30%) compared to when using the normalization preprocessing methods before the NW-D method. It can also be found by comparing the same scenarios that using complementary preprocessing methods does not always yield better correct clustering results than using a single method (NW-D+PQN → 38% Vs NW-D → 47%). This is an important consideration when using more than one preprocessing method.

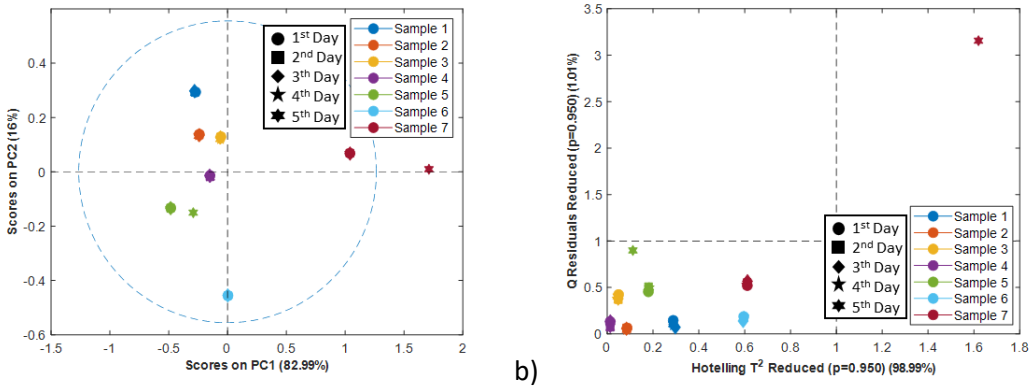
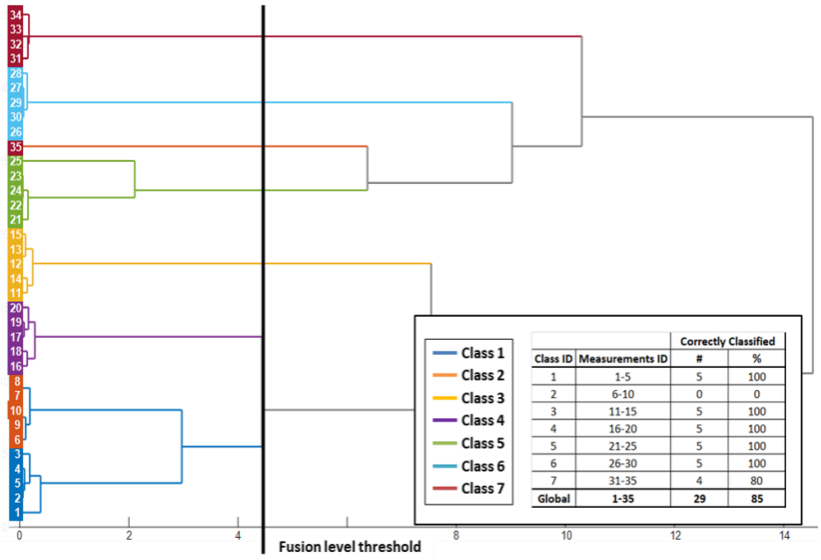


Figure 5 – a) Score plot of PC1 & PC2 for Case 1 with EMSC preprocessing, b) Reduced Qresidual & Hotelling's T^2 for Case 1 using 2 PCs with EMSC preprocessing. Legend: Shapes → day of measurement. Color → samples measured (classes)

The results obtained in this case show the proposed discrimination methodology's ability to evaluate the effectiveness of preprocessing methods in minimizing spectral variability in NIR measurements due to the lack of reproducibility. Moreover, it is worth mentioning that these results also illustrate the versatility of the proposed methodology for detecting potential anomalous data (outliers) caused by possible errors in spectrum acquisition.

As mentioned before, no detailed description for cases 2, 3, and 4 are presented; only their main results are shown. The detailed results of these cases are shown in the article's supplementary section.



Class ID	Measurements ID	Correctly Classified #	Correctly Classified %
1	1-5	5	100
2	6-10	0	0
3	11-15	5	100
4	16-20	5	100
5	21-25	5	100
6	26-30	5	100
7	31-35	4	80
Global	**1-35**	**29**	**85**

Figure 6 – Dendrogram for Case 1 (all measures) with EMSC preprocessing. Legend: Color samples measured (classes). Table: correct grouping percentage for each class

3.2 Case 2 – Repeatability

Figure 7 shows the raw spectra used in the analysis of the variability caused by repeatability measurements. Analogous to case 1, the spectral variability generated by the physicochemical nature of the sample can be observed, presenting the same relationship (trend) with the properties reported in Table 1. However, the black square in Figure 7 shows that the variability caused by repeatability measurements seems to present a similar behavior to the variability generated by the combination of two different effects (additive and multiplicative), making the spectral differences of a single sample more evident in comparison with case 1.

In this second case, a total of 21 spectra (7 samples performed in triplicate) were analyzed to demonstrate the proposed methodology's ability to evaluate the effectiveness of preprocessing methods in minimizing unwanted spectral variability due to the lack of repeatability. The correct grouping percentage achieved from all the preprocessing methods scenarios is shown in Figure App 1 (squared shape – orange color).

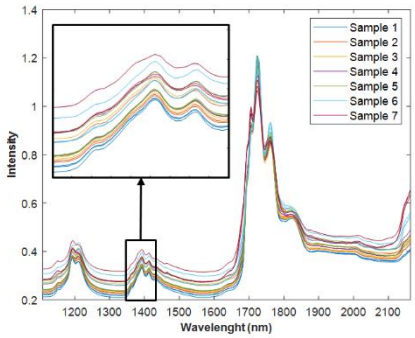


Figure 7 – 21 raw spectra used in Case 2 evaluation over the entire spectral range. Legend: Color → samples analyzed. Black square → 21 raw spectra used in Case 2 evaluation magnified over the 1340nm - 1430nm spectral range

In the scenario that is assumed that the variability generated in the repeatability measurements do not have a significant impact on the grouping of sample measurements, that is, that no additional preprocessing methods are needed besides the data centering, only 48% of the measurements were correctly grouped (see Figure App 1 (squared shape – orange color) mean center scenario). Therefore, it can be preliminarily concluded that preprocessing methods to reduce spectral variations due to repeatability are needed. From the PCA analysis (data not shown) of this scenario (centered data), it could be determined that the first repetition of each sample differs significantly from its second and third repetitions. However, all of them were within the Qresidual and Hotelling's T^2 test thresholds (data not shown). Therefore, they cannot be considered as anomalous data (outliers).

From all the evaluated scenarios, fourteen achieved a correct grouping of 100% (AWLS-B+MSC, AWLS-B+SNV, Dtd+MSC, Dtd+PQN, Dtd+SNV, EMSC+PQN, EMSC+SNV, MSC+AWLS-B, MSC+Dtd, PQN+EMSC, SNV+AWLS-B, SNV+Dtd, SNV+EMSC, and VSN). Among these scenarios, only one uses a single method, which is VSN (see HCA dendrogram in Figure App 4). Therefore, VSN could be selected as the most effective preprocessing method to minimize the spectral variability due to the lack of repeatability. As in Case 1, Figure App 1 shows that the order of use of the preprocessing methods affects the correct grouping result. Comparing the same scenarios as in Case 1, it can be reaffirmed that the results are more promising when the filtering methods are applied after the normalization methods.

From these results, it can be concluded that spectral variability due to repeatability has a lesser impact than those generated by reproducibility. Nevertheless, the proposed methodology demonstrated that the two cases' variability could be entirely compensated using an appropriate data preprocessing strategy.

3.3 Case 3 – Temperature effect

As previously mentioned in the introduction of the manuscript, sample temperature is one of the factors having significant impact on the NIR spectra acquisition. Figure 8 shows that the spectral variability generated by the sample temperature presents a behavior similar to the multiplicative effect. However due to the absorbance shift caused by the temperature increase, which prevents having a direct relationship between this parameter and the height of the acquired spectra, the spectral difference presents a nonlinear growth[45]. This can be corroborated in the black square of Figure 8, where it is observed that with a temperature variation greater than 15°C the spectral variability is more evident than when the delta in temperature is less than 15°C. This non-linear impact of the sample temperature on the spectrum acquisition could limit the performance of the different preprocessing methods evaluated.

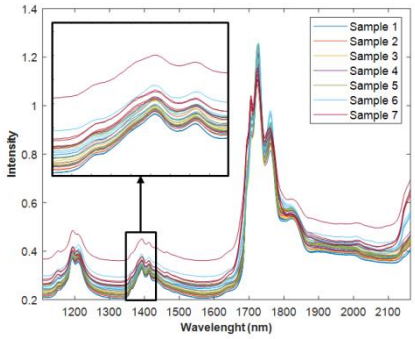


Figure 8 – a) 35 raw spectra used in Case 3 evaluation over the entire spectral range. Legend: Color → samples analyzed. Black square → 35 raw spectra used in Case 3 evaluation magnified over the 1340nm - 1430nm spectral range

In this third case, spectra acquired at 5 different sample temperatures (35 spectra) were analyzed to find the most effective preprocessing scenario to reduce undesired spectral variability due to the sample temperature. The correct grouping percentage achieved from all the evaluated preprocessing strategies is shown in Figure App 1 (triangle shape – green color).

From Figure App 1, it can be seen that if no other preprocessing method than data centering is applied, the percentage of correctly grouped measurements is 49%. This result reflects, as expected, the need to apply preprocessing methods to reduce variability caused by sample temperature. The clustering results shown in Figure App 1 reveal that no evaluated preprocessing scenario could entirely compensate the spectral variability due to temperature variations, meaning that the entire accurate measurement grouping was not achieved in any scenario. The best performing scenario is the SG-D+SNV with an accurate grouping percentage of 80%. Although Figure App 1 shows that the best performance scenario does not achieve the correct grouping of all 35 measurements, the results shown in Figure App 5 (scenario SG-D+SNV) show that samples 1, 2, 3, 5, and 6 have an accurate

grouping in all their measurements (5 out of 5 = 100%). The class affecting the overall measurement clustering is sample 4 (purple), which does not have any measurements grouped correctly. As in Case 1, it could be assumed that the whole misgrouping of sample 4 is due to the presence of some atypical data. However, no measurement was found above the Qresidual and Hotelling's T^2 tests thresholds in any scenario (data not shown).

The results analyzed in this case show that sample temperature is a very influential parameter on the spectrum acquisition. Therefore, it is recommended to use a strategy that evaluates this variable's impact more thoroughly for a more efficient solution[9, 45].

3.4 Case 4 – cases 1, 2 & 3 combined

The parameters causing unwanted spectral variability evaluated in the 3 cases previously described are likely to occur simultaneously, mainly when online NIR measurement is used for real-time data analysis. For this reason, a fourth case was evaluated where the spectral variability generated by these three cases was combined.

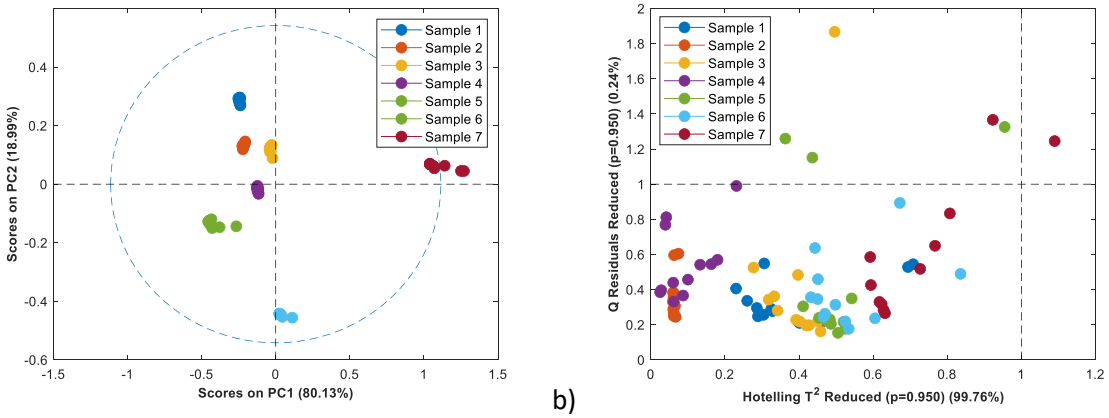
The dataset used in this case comprises 84 spectra (35 from reproducibility Case, 21 from repeatability Case, and 28 from sample temperature Case). As in the 3 cases already studied, the Qresidual and Hotelling's T^2 analyses were applied to the centered dataset to determine the possible presence of atypical data (data not shown). The measurement identified as an outlier in Case 1 (Sample 7, 5th acquisition day) once again exceeds the thresholds of the two tests. Thus, this measurement was removed, and the methodology proposed in this study was applied to the other 83 measurements.

In this case, 5 scenarios (EMSC, EMSC+PQN, PQN+EMSC, SNV+EMSC, and VSN+EMSC) had the best performance, where 80 out of 83 measurements were correctly grouped, that is, 96% of correct grouping. All 5 scenarios identified involve the use of EMSC, which, when evaluated individually, yields the same percentage of correct grouping (96%). It could be preliminarily inferred that for this case, the normalization methods do not give any additional improvement, and thus EMSC would be selected as the most effective preprocessing method scheme. Nevertheless, the use of the proposed methodology allows a more detailed analysis to determine at what level each type of variability is minimized in each of the 5 scenarios mentioned. In this way, the most effective preprocessing scheme's determination can be done more conveniently according to the researcher's objective.

Figure 9 shows the PCA score plot and the Qresidual & Hotelling's T^2 results for the VSN+EMSC preprocessing method scenario. The three measurements that were not correctly grouped correspond to the measurement at 80°C of samples 3 (highest value of residual Q test) and 4, and the

measurement at 75°C of sample 4. From Figure App 1, it is also concluded that the EMSC and VSN methods have a better performance when grouping the samples measured at different temperatures in Case 4 (circle shape – purple color, combined cases) than in Case 3 (triangle shape – green color, sample temperature case).

Although the conclusion made previously that no method can entirely compensate the measurement variability caused by sample temperature is reaffirmed, the reduction of this variability was quite considerable in delta temperature lower than 20°C.



a) b)

Figure 9 – a) Score plot of PC1 & PC2 for Case 4 with VSN+EMSC preprocessing, b) Reduced Qresidual & Hotelling's T^2 for Case 4 using 3 PCs with VSN+EMSC preprocessing. Legend: Color → samples measured (classes)

To sum up the results obtained by using the analysis methodology proposed, it could be concluded that for cases 1 and 2 (reproducibility and repeatability), the variability affecting the measurement clustering was fully compensated using a single preprocessing method. On the other hand, cases 3 and 4 (sample temperature and combination of cases) needed the combination of two methods, and still, the correct grouping of all the measurements was not achieved. The sample temperature has a high impact on the spectrum acquisition. Therefore, it is recommended to use a methodology that evaluates this variable more thoroughly for a more efficient solution [9, 45].

4 Results comparison

In order to validate the consistency and reliability of the proposed approach, the analysis methodology results were compared with those obtained by Wilks' lambda criterion.

Table 3 summarizes each case's most relevant results achieved by both the proposed approach and Wilks' Lambda.

Table 3 % Results comparison using Wilk's Lambda algorithm

Case	Preprocessing method	Methodology (% Grouped/100)	Wilk's Lambda
Case 1 (All Measurements)	Mean Center	0.46	0.95
	EMSC	0.85	0.96
Case 1 (Measurement 35 removed)	Mean Center	0.53	0.68
	EMSC	1.00	0.99
Case 2	Mean Center	0.48	0.69
	VSN	1.00	0.99
Case 3	Mean Center	0.49	0.55
	SavGol + SNV	0.82	0.95
Case 4	Mean Center	0.51	0.66
	VSN+EMSC	0.96	0.99

The results shown in

Table 3 validate the approach proposed in this study. It can be seen that the results between the two methodologies are comparable, except for Case 1, including all measurements, when the data have been only mean-centered. In this case, Wilks' Lambda value is close to 1, while the value obtained by the proposed methodology is 0.46. The difference may be attributable to the presence of the outlier identified in Case 1 and how each approach handles this type of data. While the Wilks' lambda criterion assumes that there is no presence of outliers in the analyzed dataset, the methodology used in this study provides a preliminary analysis of the dataset for the identification and removal of possible anomalous data. This premise can be supported by observing that the two approaches' results are comparable when the identified outlier is removed from the dataset (see

Table 3 - Case 1 (Measurement removed)).

Moreover, the proposed methodology provides a more detailed discrimination analysis in comparison with Wilks' lambda criterion. As a way of example, both the proposed approach and the Wilks' lambda results obtained from the individual evaluation of the nine preprocessing methods in Case 1 were compared. From Table 4, it can be seen that Wilks' lambda criterion presents no significant differences in 4 preprocessing methods (SG-D, EMSC, MSC, SNV), which could lead to the conclusion that the 4 methods have equal effectiveness in minimizing the unwanted spectral

variability. On the contrary, the proposed methodology allows to identify that out of these 4 methods, two are equally effective in minimizing the unwanted spectral variability (MSC, SNV), the least effective is the SG-D, and the most effective preprocessing method, which achieves the maximum compensation of the unwanted spectral variability, is the EMSC. Therefore, the methodology used in this work could help to select the preprocessing method in a more precise and reliable way. Finally, the proposed methodology allows identifying measurements and samples that have been properly or poorly discriminated, information that the Wilks' lambda does not provide as it is a global measurement.

Table 4 Case 1 detailed results comparison

Preprocessing Method	Explored Methodology (% Grouped/100)	Wilk's Lambda
Mean Center	0.53	0.68
AWLS-B	0.65	0.88
NW-D	0.56	0.73
SG-D	0.65	0.98
Dtd	0.65	0.90
EMSC	1.00	0.99
MSC	0.82	0.99
PQN	0.68	0.87
SNV	0.82	0.99
VSN	0.94	0.85

5 Conclusions

The results obtained in this study shows the capacity of the analysis methodology used to assess the effectiveness of preprocessing methods in reducing the undesired spectral variability of near-infrared spectroscopy (NIRS) measurements in a more thoughtfully and detailed manner than other approaches based on the dataset variance analysis such as the Wilks' lambda criterion.

In this study, an original strategy not previously reported in literature was proposed to evaluate and determine the effectiveness of different preprocessing methods in minimizing the unwanted spectral variability due to parameters related to the continuous and repetitive NIR spectra acquisition such as repeatability, reproducibility, sample temperature, and the combination of these three parameters.

It is essential to stress the twofold benefit of using the proposed methodology. On the one hand, the detailed discrimination analysis provides a significant aid in determining the most effective data preprocessing scheme. On the other hand, the methodology provides a preliminary data analysis

step for identifying and removing the potential anomalous data from the dataset, thus improving the reliability of the final results.

The results obtained using the proposed analysis methodology suggest that the variability caused by repeatability and reproducibility can be fully corrected when using the adequate preprocessing scheme; however, no preprocessing scenario could entirely compensate the unwanted spectral variability caused by the sample temperature. Similarly, the detailed discriminant analysis employed in this study showed that the EMSC preprocessing method presents interesting and promising results in all cases.

The preprocessing scheme's ultimate selection should be conducted in a careful manner considering the researcher's objective. The proposed methodology offers an analysis strategy that could help determine the most effective preprocessing scheme more reliably.

The conclusions reached in this work promote further optimization and automation of the proposed methodology to improve its implementation in large datasets.

The strategy proposed was shown to work for a case study including seven different hydrocarbon samples but can be generally applicable in any study involving spectroscopic information analysis

Acknowledgments

The authors would like to thank IFP Energie Nouvelles for providing the hydrocarbon samples obtained in their HCK pilot plant reactors and the facilities for spectra acquisition and data analysis. Thanks also go to Axel One for providing the spectrometer for the NIR spectra acquisition.

6 Declaration of conflicting interests

The Author(s) declare(s) that there is no conflict of interest

References

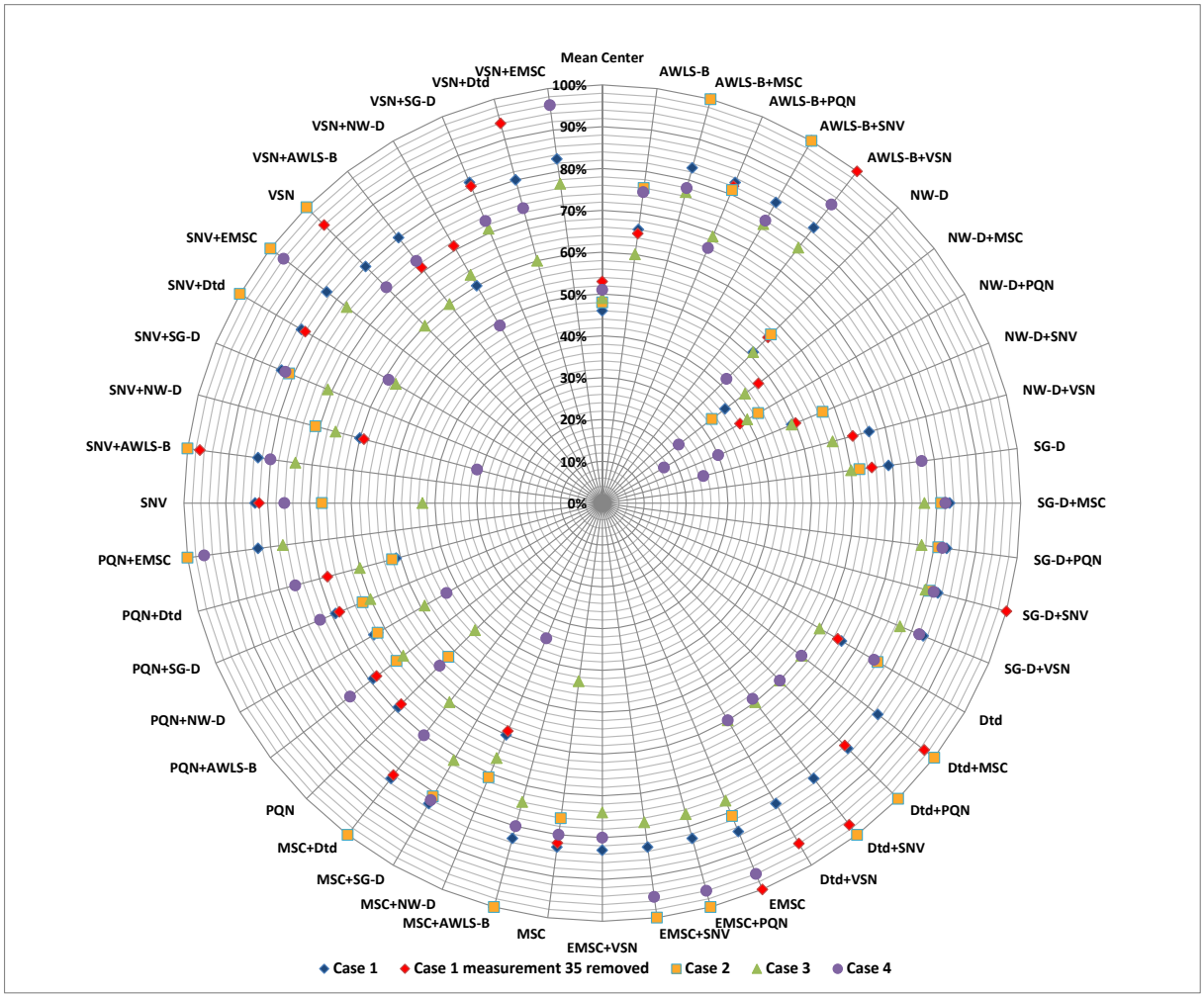
1. J. Fernández-Novales, M.-I. López, M.-T. Sánchez, J.-A. García and J. Morales, “A feasibility study on the use of a miniature fiber optic NIR spectrometer for the prediction of volumic mass and reducing sugars in white wine fermentations”, *Journal of Food Engineering*, **89**, 3 (2008).
2. W. Saeys, N. Nguyen Do Trong, R. van Beers and B.M. Nicolaï, “Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: A review”, *Postharvest Biology and Technology*, **158** (2019).
3. M. Blanco and A. Peguero, “Analysis of pharmaceuticals by NIR spectroscopy without a reference method”, *TrAC Trends in Analytical Chemistry*, **29**, 10 (2010).

4. T. de Beer, A. Burggraeve, M. Fonteyne, L. Saerens, J.P. Remon and C. Vervaet, “Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes”, *International journal of pharmaceutics*, **417**, 1-2 (2011).
5. R.M. Balabin, E.I. Lomakina and R.Z. Safieva, “Neural network (ANN) approach to biodiesel analysis: Analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy”, *Fuel*, **90**, 5 (2011).
6. Zanier-Szydlowski N., Quignard A., Baco F., Biguerd H., Carpot L., “Control of Refining Processes on Mid-Distillates by Near Infrared Spectroscopy”, *Oil & Gas Science and Technology - Rev. IFP* (1999).
7. P.R. Wahl, I. Pucher, O. Scheibelhofer, M. Kerschhaggl, S. Sacher and J.G. Khinast, “Continuous monitoring of API content, API distribution and crushing strength after tableting via near-infrared chemical imaging”, *International journal of pharmaceutics*, **518**, 1-2 (2017).
8. R.R. de Oliveira, R.H.P. Pedroza, A.O. Sousa, K.M.G. Lima and A. de Juan, “Process modeling and control applied to real-time monitoring of distillation processes by near-infrared spectroscopy”, *Analytica chimica acta*, **985** (2017).
9. F. Chauchard, J.M. Roger and V. Bellon-Maurel, “Correction of the temperature effect on near infrared calibration—application to soluble solid content prediction”, *Journal of Near Infrared Spectroscopy*, **12** (2004).
10. B. Igne, M.N. Hossain, J.K. Drennen and C.A. Anderson, “Robustness Considerations and Effects of Moisture Variations on near Infrared Method Performance for Solid Dosage Form Assay”, *Journal of Near Infrared Spectroscopy*, **22**, 3 (2014).
11. J.M. Betz, P.N. Brown and M.C. Roman, “Accuracy, precision, and reliability of chemical measurements in natural products research”, *Fitoterapia*, **82**, 1 (2011).
12. A.C. Olivieri and N.M. Faber, “Validation and Error”, in: *Comprehensive chemometrics. Chemical and biochemical data analysis,* Ed by S.D. Brown, Elsevier, Amsterdam, pp. 91–120 (2009).
13. ISO, “Measurement management systems - Requirements for measurement processes and measuring equipment”, 10012:2003. UNE.
14. Florian Wülfert,†, Wim Th. Kok,† and, and Age K. Smilde*,†, “Influence of Temperature on Vibrational Spectra and Consequences for the Predictive Ability of Multivariate Models”, *Analytical Chemistry*, **70** (1998).
15. Hideyuki Abe, Chie Iyo, and Sumio Kawano, “A Study on the Universality of a Calibration with Sample Temperature Compensation”, *Journal of Near Infrared Spectroscopy*, **8** (2000).
16. W.G. Hansen, S.C.C. Wiedemann, M. Snieder, and V.A.L. Wortel, “Tolerance of near Infrared Calibrations to Temperature Variations; A Practical Evaluation”, *Journal of Near Infrared Spectroscopy*, **8** (2000).
17. Jasem M. Al-Besharah/Saed A. Akashah/Clive J. Mumford, “The effect of temperature and pressure on the viscosities of crude oils and their mixtures”, *Industrial & Engineering Chemistry* (1989).
18. P. Luo and Y. Gu, “Effects of asphaltene content on the heavy oil viscosity at different temperatures”, *Fuel*, **86**, 7-8 (2007).
19. R. Payri, F.J. Salvador, J. Gimeno and G. Bracho, “The effect of temperature and pressure on thermodynamic properties of diesel and biodiesel fuels”, *Fuel*, **90**, 3 (2011).
20. Å. Rinnan, F. van den Berg and S.B. Engelsen, “Review of the most common pre-processing techniques for near-infrared spectra”, *TrAC Trends in Analytical Chemistry*, **28**, 10 (2009).
21. Abraham. Savitzky/M. J. E. Golay, “Smoothing and Differentiation of Data by Simplified Least Squares Procedures.”, *Analytical Chemistry*, **36** (1964).

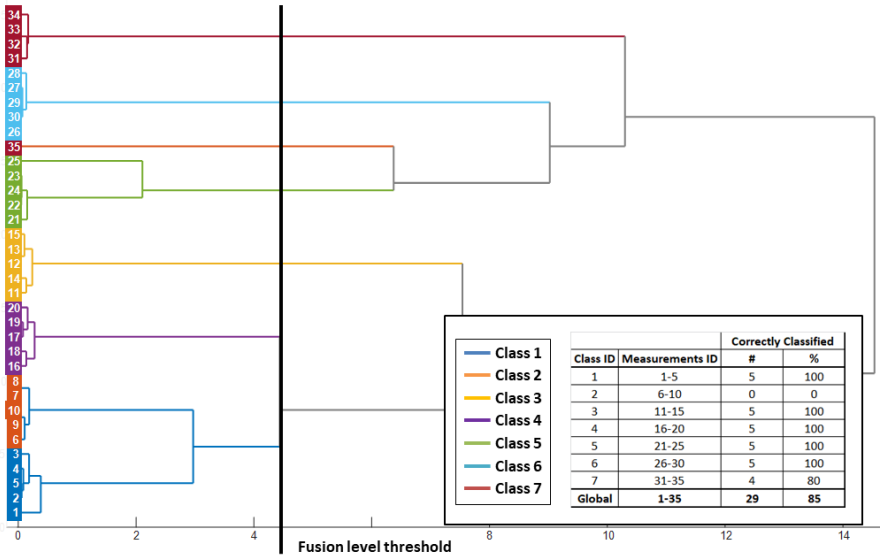
22. M. Harald and Edward Stark, “Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy”, *Journal of Pharmaceutical & Biomedical Analysis*, **9** (1991).
23. Barnes, R. J., Dhanoa, M. S., & Lister, S. J., “Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra.”, *Applied Spectroscopy*, **43** (1989).
24. G. Rabatel, F. Marini, B. Walczak and J.-M. Roger, “VSN: Variable sorting for normalization”, *Journal of Chemometrics*, **34**, 2 (2020).
25. A. Gholizadeh, L. Borůvka, M.M. Saberioon, J. Kozák, R. Vašát and K. Němeček, “Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features”, *Soil & Water Res.*, **10**, No. 4 (2016).
26. Y. Jiao, Z. Li, X. Chen and S. Fei, “Preprocessing methods for near-infrared spectrum calibration”, *Journal of Chemometrics*, **130**, 1 (2020).
27. C.D. Brown, L. Vega-Montoto and P.D. Wentzell, “Derivative Preprocessing and Optimal Corrections for Baseline Drift in Multivariate Calibration”, *Appl Spectrosc*, **54**, 7 (2000).
28. J. Gerretzen, E. Szymańska, J.J. Jansen, J. Bart, H.-J. van Manen, E.R. van den Heuvel and L.M.C. Buydens, “Simple and Effective Way for Data Preprocessing Selection Based on Design of Experiments”, *Analytical Chemistry*, **87**, 24 (2015).
29. O. Devos and L. Duponchel, “Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression”, *Chemometrics and Intelligent Laboratory Systems*, **107**, 1 (2011).
30. F. Allegrini and A.C. Olivieri, “An integrated approach to the simultaneous selection of variables, mathematical pre-processing and calibration samples in partial least-squares multivariate calibration”, *Talanta*, **115** (2013).
31. D.L. Luthria, S. Mukhopadhyay, L.-Z. Lin and J.M. Harnly, “A comparison of analytical and data preprocessing methods for spectral fingerprinting”, *Applied Spectroscopy*, **65**, 3 (2011).
32. S. S Wilks, “The large-sample distribution of the likehoold ratio for testing composite hypotheses”, *Mathematical Statistics* (1962).
33. L. Rokach and O. Maimon, “Clustering Methods”. Springer-Verlag.
34. Maureen Bricker, Vasant Thakkar, John Petri, “Hydrocracking in Petroleum Processing”, in: *Handbook of Petroleum Processing 2014*, pp. 1–35.
35. J.G. Speight, “Hydrocracking”, in: *The Refinery of the Future,* Elsevier, pp. 275–313 (2011).
36. ASTM D1218 - 12, “Standard Test Method for Refractive Index and Refractive Dispersion of Hydrocarbon Liquids”, https://www.astm.org/Standards/D1218.htm.
37. ASTM D2887 - 19ae1, “Standard Test Method for Boiling Range Distribution of Petroleum Fractions by Gas Chromatography”, https://www.astm.org/Standards/D2887.htm.
38. K.H. NORRIS and P.C. WILLIAMS, “Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat. I. Influence of particle size.”, *Cereal Chemistry*, **61** (Mar. 1984).
39. H. Martens and T. Naes, *Multivariate calibration*. Wiley, Chichester (1989).
40. F. Dieterle, A. Ross, G. Schlotterbeck and H. Senn, “Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics”, *Analytical Chemistry*, **78**, 13 (2006).
41. S. Wold, K. Esbensen and P. Geladi, “Principal component analysis”, *Chemometrics and Intelligent Laboratory Systems*, **2**, 1-3 (1987).

589 42. L.S. Chen, D. Paul, R.L. Prentice and P. Wang, “A regularized Hotelling's T2 test for pathway
590 analysis in proteomic studies”, *Journal of the American Statistical Association*, **106**, 496 (2011).
591 43. K. Sasirekha and P. Baby, “Agglomerative Hierarchical Clustering Algorithm- A Review”,
592 *International Journal of Scientific and Research Publications*, **3** (2013).
593 44. F. Murtagh and P. Legendre, “Ward’s Hierarchical Agglomerative Clustering Method: Which
594 Algorithms Implement Ward’s Criterion?”, *J Classif*, **31**, 3 (2014).
595 45. J.-M. Roger, F. Chauchard and V. Bellon-Maurel, “EPO–PLS external parameter orthogonalisation
596 of PLS application to temperature-independent measurement of sugar content of intact fruits”,
597 *Chemometrics and Intelligent Laboratory Systems*, **66**, 2 (2003).
598

599 # Appendix
600



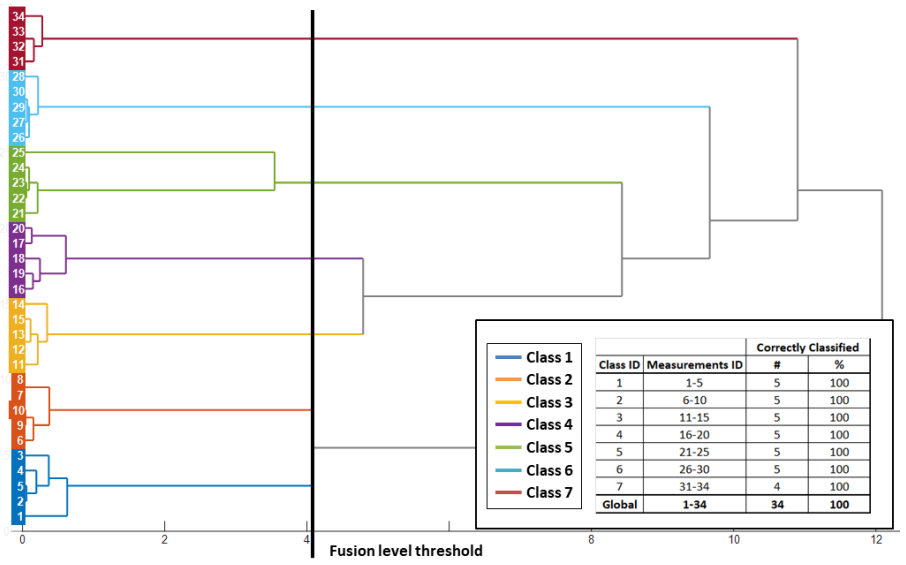
601
602 **Figure App 1 Comparative summary of the effectiveness of the preprocessing scenarios evaluated in the 4 case studies.**
603 **Legend : Shapes and color → Case studies. Description : Circumferential lines → Percentage of correct grouping (0%**
604 **Center - 100% outer line). Radial lines → Evaluated preprocessing scenario, from left to right the order of use of the**
605 **methods.**



Class ID	Measurements ID	Correctly Classified #	Correctly Classified %
1	1-5	5	100
2	6-10	0	0
3	11-15	5	100
4	16-20	5	100
5	21-25	5	100
6	26-30	5	100
7	31-35	4	80
Global	**1-35**	**29**	**85**

606
607 **Figure App 2 Dendrogram for Case 1 (all measurements) with EMSC preprocessing. Legend: Color samples measured**
608 **(classes). Table: correct grouping percentage for each class**

609



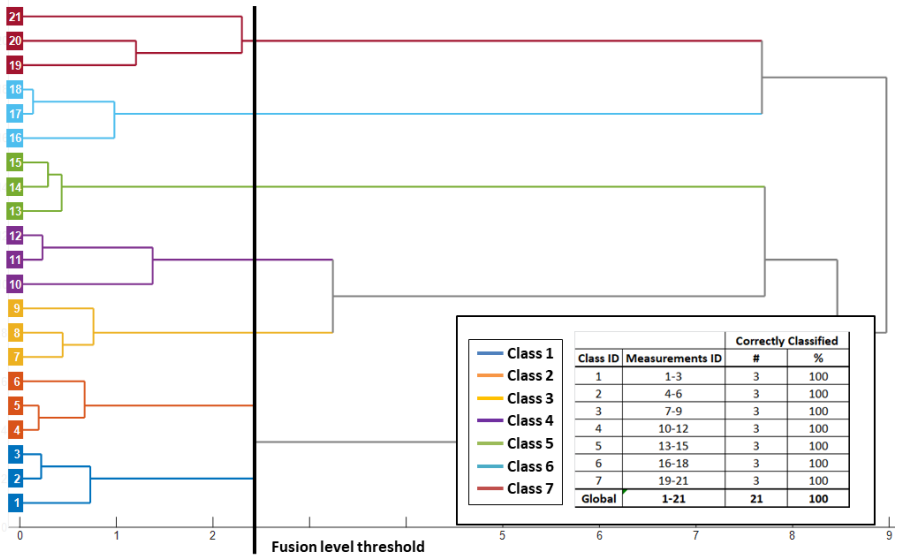
Class ID	Measurements ID	Correctly Classified #	Correctly Classified %
1	1-5	5	100
2	6-10	5	100
3	11-15	5	100
4	16-20	5	100
5	21-25	5	100
6	26-30	5	100
7	31-34	4	100
Global	**1-34**	**34**	**100**

610

611 **Figure App 3 Dendrogram for Case 1 (removing measurement 35) with EMSC preprocessing. Legend: Color samples**
612 **measured (classes). Table: correct grouping percentage for each class**

613

614

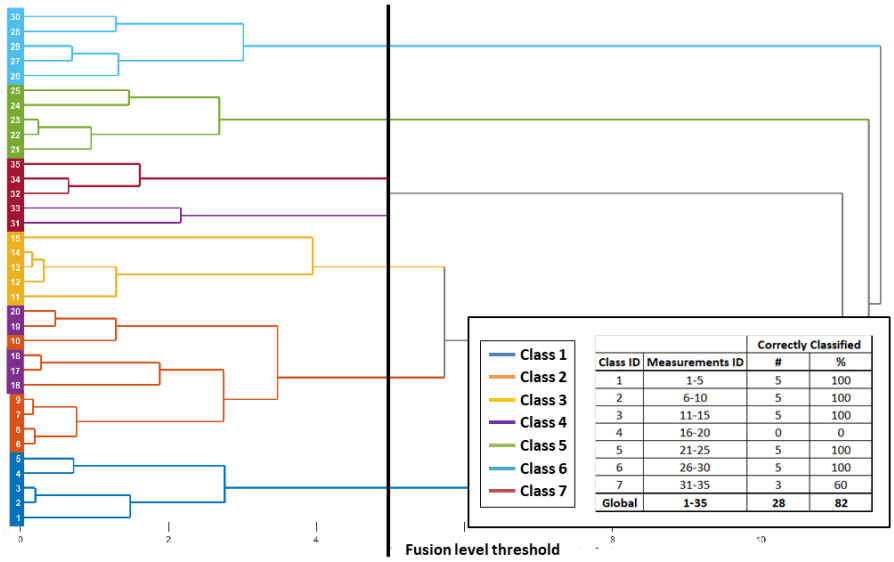


Class ID	Measurements ID	Correctly Classified #	Correctly Classified %
1	1-3	3	100
2	4-6	3	100
3	7-9	3	100
4	10-12	3	100
5	13-15	3	100
6	16-18	3	100
7	19-21	3	100
Global	1-21	21	100

615

616 **Figure App 4 dendrogram for Case 2 with VSN preprocessing. Legend: Color samples measured (classes). Table:** correct
617 **grouping** percentage **for each class**

618

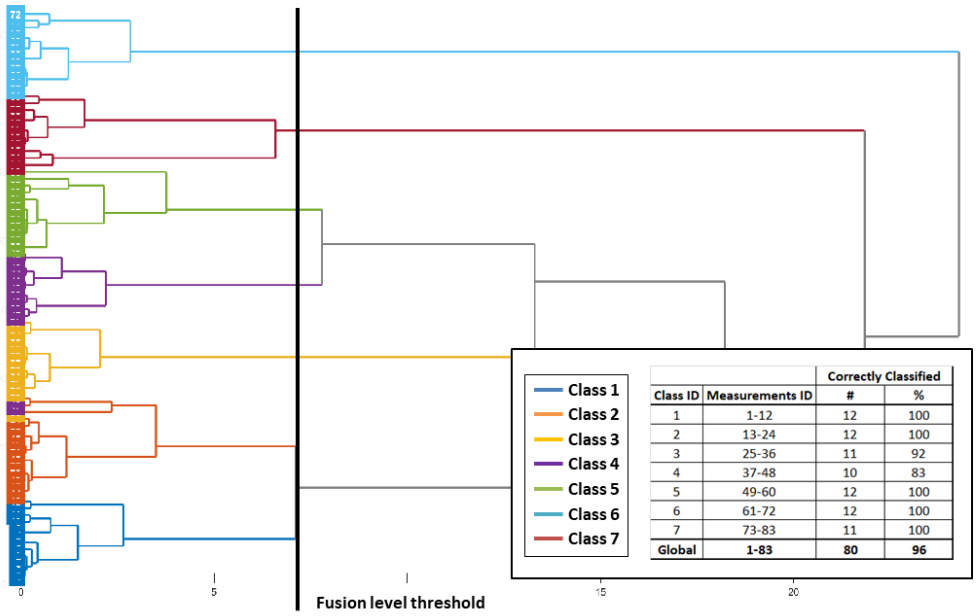


Class ID	Measurements ID	Correctly Classified #	Correctly Classified %
1	1-5	5	100
2	6-10	5	100
3	11-15	5	100
4	16-20	0	0
5	21-25	5	100
6	26-30	5	100
7	31-35	3	60
Global	1-35	28	82

619

620 **Figure App 5 dendrogram for Case 3 with SG-D+SNV preprocessing. Legend: Color samples measured (classes). Table:**
621 **correct grouping percentage for each class**

622



		Correctly Classified	
Class ID	Measurements ID	#	%
1	1-12	12	100
2	13-24	12	100
3	25-36	11	92
4	37-48	10	83
5	49-60	12	100
6	61-72	12	100
7	73-83	11	100
Global	**1-83**	**80**	**96**

Figure App 6 dendrogram for Case 4 with VSN+EMSC preprocessing. Legend: Color samples measured (classes). Table: correct grouping percentage for each class